社会人のための確率・統計—試論

藤 曲 哲 郎

1．はじめに

大学や各種研究所の理工系，生命科学系，行動科学系，社会科学系等々の研究者はそれぞれの研究上の目的から，ある実験なり調査なりを行い，得られた実験データや調査結果を集計して，パラメータの推定をしたり回帰分析をしたり予測をしたり，集団間の差異を判定したり理論上の仮説を統計的仮説検定法を用いて検証したりする。これらの各種方法が総じて統計学を構成し，それぞれ統計学の基礎理論に基づいて開発されている。

このように研究者にとって統計学の基礎的理解は不可欠であることは周知であると思われるが，実は一般の社会人にとってもそのことは不可欠ではないにしても大切である。確かに大部分の社会人は自らデータの収集やそれに基づく統計的推測を行うことはないであろう。ところが今日実際には新聞やテレビの報道でも連日のように何らかの統計的数値が報じられている。現内閣の支持率，人口予測，各種経済指数などは代表的なものであろう。それらの多くは国の統計局や報道機関を含む各種調査機関によってなされた調査結果の報告と，それに基づく論評などである。それは数値だけの場合もあるが，しばしば棒グラフや円グラフなどを利用して示されている。一般の社会人はこれらの調査結果をどのように理解しているのだろうか。

統計数値にも大きく分ければ2種類あって，一つは何年何月何日現在の日本の人口は幾人であったとか，何年度の日本の交通事故による死者数は何人であったとか言う単に事実を示す数値であるが，もう一つは調査結果を基に推定された数値である。上の現内閣の支持率等の大部分の数値は後者に属するものであって，これらは前者の数値のように確定したものでは決してないことを忘れてはならないのである。この後者の推定結果の中には医療関係の定期検査が本当に有効かどうかの判定も含まれているし，新薬がその効能を本当に期待できるかどうかも含まれている。短期あるいは長期の経済予測も含まれ，これらは多くの政策決定に影響している。その他にもテレビ番組の視聴率，子供の学力の偏差値に基づく合否の予測，天気予報での降水確率，選挙の出口調査による当選者予測等々がある。

このように多くの統計的推測結果が我々の生活，社会の動向に強く関っている。これらの統計数値が推測であって単に確定した数値ではない，という事実をどれほど多くの人々が承知しているであろうか。正しい判断をするためにも，判断を誤らないためにも，あるいは誤った判断をしてしまう危険を最小限にするためにも，社会人の誰にとっても推測統計学の初歩的理解は必要であろう。

筆者：理学部教授
受理：　平成9年2月25日

推測統計学は確率論をその理論的基礎にしている。そのため確率論の基礎的事項にたいする初歩的理解は必要になる。他方，偶然の影響を避けられないような現象を理解するには確率を用いる必要があるので，確率に関る知識は広く応用される。確率は大学でも統計学の講義でその基礎として始めに教えられることが多いが，実際には学生はそれを初歩的にも十分には理解できないでいることが多いといえよう。そのため，後で教えられる各種統計的方法もそれぞれ単なるノウハウとしてしか学習できない。本当は，確率を十分に理解できれば，推測統計学の各種方法の考え方を理解することは困難ではないはずである。

ここでは確率に関する各種重要事項を大学生に限らず一般社会人を含む高校生以上の誰にでも理解できるように，コンピュータによるシミュレーションを活用する方法について試論を述べる。この特徴は幾何学の理解のために実際の図形を活用するように，確率・統計の「理解」のためにコンピュータ シミュレーションを活用することにあって，データ処理などの「実務」をコンピュータにやらせるための方法ではない。これは筆者が金沢大学の主として１年生を対象にしている授業を元に，社会人のために金沢大学教育開放センターの一講座として，３時間ごと５回にわたって金沢大学情報処理センターの施設を利用して行った経験を踏まえたものである。この講座の実施に当たって，励ましとアドバイスをいただいた佐伯開放センター長と講座の助手をしてくれた大学院生の示村，金両君および熱心に参加してくれた受講生達に感謝する。

第２節で取り上げるコンピュータ プログラムの多くは筆者など共著の参考文献［1］から引用した。富士通製のF-BASIC/WinというWindows 3.1の上で動作するBASIC用に書かれたもので，他のBASICとはLINE文のパラメータの書き方などで多少異なる点があるが，殆どは変わらないはずである。ここではプログラムの各命例文についての説明は省略する。また，実際に開放講座で利用したプログラムより一部分簡略化した。

２．計算機実験による確率

確率は数学的には全測度が１の正の測度として定義され，確率論は完全に数学の一分野であるが，統計学その他への応用を考える場合はそれだけの理解では十分とはいえない。特に始めて確率について学習し，さらにそれを応用したい場合は日常的に使われている「確率」についての考察を欠かせない。理論上の確率と統計学等の実際に用いられる確率の応用との関係を理解するためには，一つの実験を同一条件の下で極めて多数回繰り返して結果を観察しなければならない。あるいはそれと同等のデータが得られるような実験，観測等をする必要がある。簡単なコイン投げやサイコロ投げの実験でもそれらを例えば１万回繰り返すのは実際的ではないが，このような実験であれば簡単なプログラムを書いてパソコンで実行することが出来る。ただし，その際パソコンは乱数を利用するが，その乱数自身をパソコンは計算する。従って，本来の意味での乱数でないことは明白で擬似乱数と言われることになる。

(1) 擬似乱数　パソコンでどのように乱数が発生されるかを知るために，次の２つのプログラムを実行する。

[2-1-1.BAS]

```
100 RANDOMIZE TIMER
110 FOR I=1 TO 100
120  PRINT RND,
130 NEXT I
```

この2-1-1.BASを実行することで，パソコンの画面上に100個の擬似乱数が表示される。これらは0と1の間のデタラメな数値であるはずであるが，実際には一定の仕方で計算されたものである。これを理解するために次のプログラムを実行する。

[2-1-2.BAS]

```
100 A=7 : B=3 : C=2^(11)-1 : K=4
110 X=K
120 FOR I=1 TO 100
130  Y=A*X+B : X=Y MOD C
140  PRINT X/C;
150 NEXT I
```

この結果，画面には0と1の間のデタラメにみえる100個の数値が表示される。このような線形合同法と呼ばれる簡単な方法でも乱数らしき数値を発生できることを知る。

(2) コイン投げのシミュレーション　上のようにパソコンで発生される乱数は本来の乱数ではないことを承知の上で，この乱数を利用して簡単なコイン投げの実験をする。

[2-2-1.BAS]

```
100 RANDOMIZE TIMER
110 INPUT P
120 FOR I=1 TO 100
130  IF RND<P THEN C=1 ELSE C=0
140  PRINT C;
150 NEXT I
```

これを実行して表の出る確率Pに0と1の間の数値を適当に入力すると，そのようなコインを100回繰り返し投げた結果が表を1，裏を0として画面に表示される。この実験を幾度か繰り返しながら画面に表示される結果を観察すると，いろいろなことに気付く。例えば，意外に0または1が続いて出ることが多いことなど。表の出る確率Pを変えて，更に実験を繰り返せば1の出方の変化にも興味が湧く。

(3) **割合と確率**　一例としてコイン投げのシミュレーションを利用して大数の法則を考える。これによって日常用いられる割合とか比率というものと理論上の確率との密接な関係を知る。

[2-3-1.BAS]

```
100 RANDOMIZE TIMER
110 INPUT P
120 INPUT N
130 FOR I=1 TO N
140   IF RND<P THEN C=1 ELSE C=0
150   PRINT C;
160   S=S+C
170 NEXT I
180 PRINT:PRINT N,S,S/N
```

このプログラムを実行して，1回のコイン投げで表の出る確率Pに例えば0.5を入力して，コインを投げる回数Nに例えば100を入力すれば，公正なコインを100回投げた結果が表は1，裏は0として画面に表示された後，最後にその実験で表が出た割合が表示される。表の出る確率を0.5に固定して，コインを投げる回数Nをいろいろ変えて，この実験を繰り返して最後に表示される表が出た割合に注意する。コインを投げる回数が100回以上であれば，殆の場合にその割合が0.5に近いことに気付くであろう。このことは次のプログラムを実行することによって明瞭となる。

[2-3-2.BAS]

```
100 N=200
110 INPUT P
120 DIM S(N):S(0)=0
130 RANDOMIZE TIMER
140 FOR I=1 TO N
150   IF RND<P THEN S(I)=S(I-1)+1 ELSE S(I)=S(I-1)
160 NEXT I
170 LINE (50,420)-(50+3*N,420),PSET,10
180 LINE (50,420)-(50,70),PSET,10
190 LINE (45,420-180)-(55,420-180),PSET,10
200 LINE (50+150,420)-(50+150,405),PSET,10
210 FOR I=1 TO N
220   PSET (50+3*I,420-INT(S(I)*180/(I*P))),9
230 NEXT I
```

このプログラムを実行して画面に表示される折れ線は，1回のコイン投げで表の出る確率Pに0と1の間の数値を適当に入力した結果，このようなコインを続けて200回まで投げたときに，各回までの表の出た割合の変動を表している。これを見れば表の出た割合が入力したPの値に近づく様子が良く分かる。このシミュレーションをPの値をいろいろに変えて繰り返すことによって，表の出る割合と1回のコイン投げで表の出る確率との関係が経験的に理解される。

こうして確率の存在を認識した上で，その一般的な特徴を明らかにする仕方で確率を公理論的に定義する。

(4) 確率変数と確率分布　次の例を考える。

例　セールスマンがある製品を売るために家庭を訪問する。この製品が売れる確率は1軒につき0.05であるという。セールスマンが1日に20軒を訪問したとき何個ぐらい売れるだろうか。

この例で1日に売れた製品の個数Xは0から20までのどれかの数であるが，それが幾つであるか予想ができない。これを次のプログラムでシミュレートする。

[2-4-1.BAS]

```
100  INPUT K
110  RANDOMIZE TIMER
120  N=20 : P=0.05
130  DIM A(N)
140  FOR I=1 TO K
150    X=0
160    FOR J=1 TO N
170      IF RND>=P THEN 190
180      X=X+1 : PRINT " O" ; : GOTO 200
190      PRINT " *" ;
200    NEXT J
210    A(X)=A(X)+1
220    PRINT X,
230  NEXT I : PRINT
240  FOR X=0 TO N
250     PRINT A(X)/K;
260  NEXT X
```

これを実行して実験回数（あるいはセールスマンが売り歩く日数）Kに例えば100を入力すると，100日間のセールスマンの実績が画面に表示される。売れたら O で売れなければ * で結果が表される。最後に1日に売れた個数Xの100日間を通しての割合が表示される。

日数Kの値を200,300,...と変えてシミュレーションを繰り返し，それらの結果に注意すれ

ば, 個数Xはランダムであるがそのそれぞれの値の割合は殆ど一定であることに気付く。こうして確率変数にはその確率分布があることを知る。

(5) 期待値と大数の法則　サイコロを1回投げて出る目の数をXとする。確率変数Xの平均を求める。

[2-5-1.BAS]

```
100  INPUT K
110  RANDOMIZE TIMER
120  DIM A(6)
130  FOR I=1 TO K
140    J=INT(6*RND)+1
150    PRINT J;
160    A(J)=A(J)+1
170    W=W+J
180  NEXT I : PRINT : PRINT
190  FOR J=1 TO 6
200    PRINT A(J)/K;
210  NEXT J : PRINT : PRINT
220  PRINT K,W/K
```

このプログラムを実行して, サイコロを投げる回数Kに例えば100を入力すると, 画面に100回サイコロを投げた結果が表示され, それぞれの目の出た割合が表示される。最後に表示されるのは100回分の出た目の平均である。回数Kを100,200,...と変えて, このシミュレーションを繰り返すことで確率変数Xの確率分布とその期待値の存在を認識できる。サイコロを投げる回数が変わる毎の平均の変動は次のプログラムによって視覚的に理解できる。

[2-5-2.BAS]

```
100  N=200
110  D=6
120  DIM X(N),S(N)
130  RANDOMIZE TIMER
140  FOR I=1 TO N
150    X(I)=INT(D*RND)+1 : S(I)=S(I-1)+X(I)
160  NEXT I
170  LINE (50,420)-(50+3*N,420),PSET,10
180  LINE (50,420)-(50,70),PSET,10
190  LINE (45,420-180)-(55,420-180),PSET,10
```

```
200  LINE (50+150,420)-(50+150,405),PSET,10
210  FOR I=1 TO N
220    PSET (50+3*I,420-INT(S(I)*2*180/(I*(D+1)))),9
230  NEXT I
```

このプログラムを繰り返し実行して画面に表示される折れ線グラフを見ることによって，出た目の平均がサイコロを投げる回数が多くなるに従って一定の値（期待値3.5）に近づく様子が良く分かる。これとプログラム2-3-2.BASとを合わせて大数の法則の存在を予見することが出来る。

(6) 分散の必要性　確率分布をその代表値としての期待値だけでは考えられない例を一つ取り上げる。

例　ある宝くじは，当たりくじには500万円，その他はすべてはずれとなっている。この当たりくじを引く確率は0.0001であるが，このことは公表されていない。400円払ってこのくじを1回引くと，平均100円の利益があるとだけ宣伝している。

このくじを引いたときの様子をシミュレーションで調べてみる。

[2-6-1.BAS]

```
100  INPUT K
110  RANDOMIZE TIMER
120  FOR I=1 TO K
130    IF RND>=0.0001 THEN 150
140    X=5000000-400 : GOTO 160
150    X=-400 : R=R+1
160    PRINT X; : Y=Y+X
170  NEXT I : PRINT : PRINT
180  PRINT 1-R/K,Y/K
```

このプログラムを実行して，くじを引く回数Kとして例えば500を入力すれば，500回のくじ引きで得られた金額がそれぞれ画面に表示される。最後に当たりくじを引いた割合と500回で得られた金額の平均が表示される。このシミュレーションを繰り返して見れば，実際には当たりくじを引くことはなく平均獲得金額は100円ではなく，400円の損失であったことに気付くだろう。500回を1000回に増やしても結果は殆ど変わらない。

この例では1回のくじ引きで得られる金額X円はランダムで理論上は確かに期待値は100である。こうして確率分布を特徴付けるもう一つの分散の必要性が導かれる。

(7) 標本平均の確率分布　大数の法則から標本平均は標本数が大きければ殆ど期待値に近い値

を取ることはすでに理解されたが，さらに詳しく標本平均の確率分布を知ることが理論だけでなく応用上も重要である。この確率分布は正規分布に近いことが知られているが，その理論(中心極限定理）を理解するのは容易ではない。そこで，サイコロ投げのシミュレーションを繰り返して，出た目の標本平均の度数分布図を描いてみる。

[2-7-1.BAS]

```
100 INPUT " Dice " ;D
110 INPUT " Trials " ;N
120 INPUT " Samples " ;K
140 DIM Y(K),D(20)
150 PRINT
160 RANDOMIZE TIMER
170 FOR I=1 TO K
180   S=0
190   FOR J=1 TO N
200     S=S+INT(D*RND)+1
210   NEXT J
220   Y(I)=SQR(3)*(2*S-N*(D+1))/SQR(N*(D*D-1))
230   Y=INT(2*Y(I)+0.5)+10
240   IF Y<0 THEN Y=0
250   IF Y>20 THEN Y=20
260   D(Y)=D(Y)+1
270   LINE (180,429)-(640,429),PSET,5
280   LINE (409,50)-(409,434),PSET,5
290   LINE (449,426)-(449,434),PSET,5
300   LINE (369,426)-(369,434),PSET,5
310   LINE (200+20*Y,429-2*D(Y))-(219+20*Y,430-2*D(Y)),PSET,3,BF,3
320 NEXT I
```

このプログラムを実行して始めに入力を求められるDice?に対して6を入力する（コインやサイコロ等で同じ実験を出来るようにしてあるので，コインの場合は2を入力する)。次のTrials?にはサイコロを投げる回数として例えば100を入力する。画面にはサイコロを100回投げた結果出た目の標本平均値に対応した小さなブロックが表示され，それはSamples?で入力した実験回数分だけ次々と積み上げられる。こうして標本平均に対応した度数分布図がダイナミックに表示される。プログラムでは標本平均そのものでなく，それを期待値が0で分散が1となるように規格化したものの度数分布を描くようになっている。理論上これは標準正規分布の密度曲線と大体一致するはずであるが，実際はどの様であるかが分かり，後で統計学などで中心極限定理を応用するときの妥当性及び限界について知ることが出来る。

3．統計的推論の計算機実験

確率論を基礎にしてどのように統計的な推論がなされるのか，またその方法及び結論は実際の応用でも妥当か，についてシミュレーションを利用して考える。

(1) 母平均の推測　大数の法則から母平均の推定値として標本平均を利用できることは，プログラム2-3-1.BAS, 2-3-2.BASによるコイン投げの実験，及び2-5-1.BAS, 2-5-2.BASによるサイコロ投げの実験によって理解できる。さらに，その推定値の誤差を明らかにするには中心極限定理が応用される。こうして導かれたのが信頼区間であるが，その信頼度の計算には極限分布である正規分布を用いている。従って，信頼度そのものも確定したものであるとは言えない。次のプログラムで信頼区間の理論で与えられた信頼度0.95が実際にあるかどうかをシミュレーションによって検討する。

[3-1-1.BAS]

```
100  INPUT " Trials " ;K
110  INPUT " Samples " ;N
120  INPUT " Expectation " ;E
130  INPUT " Standard deviation " ;S
140  PRINT
150  RANDOMIZE TIMER
160  FOR I=1 TO K
170    M=0 : V=0
180    FOR J=1 TO N
190      X=S*SQR(-2*LOG(RND))*COS(2*3.14159*RND)+E
195  REM     X=S*2*SQR(3)*RND+E-S*SQR(3)
200      M=M+X
210      V=V+X*X
220    NEXT J
230    M=M/N
240    V=V/N-M*M
250    W=1.96*SQR(V/N)
260    PRINT I; : PRINT ," Mean=" ;M;
270    PRINT ," Standard deviation=" ;SQR(V)
280    PRINT ," Confidence interval(95%)=" ; M;
290    PRINT "+-" ;W;
300    IF (E>M-W) AND (E<M+W) THEN GOTO 310 ELSE GOTO 320
310    T=T+1 : PRINT "   Yes" : GOTO 330
320    PRINT "   No!"
```

```
330  NEXT : PRINT
340  PRINT "  Rate of success=" ;T/K
```

このプログラムを実行して，Trials?に対して実験回数Kを入力し，Samples?に対しては標本数Nを入力する。続いてExpectation?に対して母平均Eを，Standard deviation?に対しては標準偏差Sの値を入力すれば，画面には期待値E，標準偏差Sの正規母集団からの大きさNの標本について計算された信頼区間が表示され，さらにそれが母平均を含んでいればYesを，含んでいなければNo!が表示される。この実験がK回繰り返されて，最後にRate of successとしてYesであった実験の割合が表示される。この割合が理論上の信頼度0.95を越えるかどうかが調べられる。母集団分布が正規分布でない場合のシミュレーションは，例えば一様分布とした場合はプログラムの190行を195行に変更することで実行できる。

このようなシミュレーションによって，理論上の信頼区間を実際にも利用できることと同時に応用上の問題点の所在を認識することができる。

(2) 相関係数の検定　2つの変量に相関があるかないかの統計的仮説検定には，理論上2次元正規母集団からの標本について，その標本相関係数を一定の仕方で変換したものが t-分布に従うことを利用する。これに対してシミュレーションを用いることによって，標本相関係数そのものについての棄却域の限界値を直接的に求められる。その限界値はシミュレーションをするごとに多少変動するが，実際の仮説検定に応用可能である。このことを次のプログラムによって調べる。

[3-2-1.BAS]

```
100  INPUT " Trials " ;K
110  INPUT " Samples " ;N
120  INPUT " Expectations:e1,e2 " ;E1,E2
130  INPUT " Standard deviations:s1,s2 " ;S1,S2
140  DIM R(K),D(30)
150  PRINT : H=4
160  RANDOMIZE TIMER
170  FOR I=1 TO K
180    M1=0 : M2=0 : V1=0 : V2=0 : C=0
190    FOR J=1 TO N
200      U1=RND : U2=RND
210      X=S1*SQR(-2*LOG(U1))*COS(2*3.14159*U2)+E1
215  REM     X=S1*2*SQR(3)*U1+E1-S1*SQR(3)
220      Y=S2*SQR(-2*LOG(U1))*SIN(2*3.14159*U2)+E2
225  REM     Y=S2*2*SQR(3)*U2+E2-S2*SQR(3)
230      M1=M1+X : M2=M2+Y
240      V1=V1+X*X : V2=V2+Y*Y
```

```
250      C=C+X*Y
260    NEXT J
270    R(I)=(N*C-M1*M2)/SQR((N*V1-M1*M1)*(N*V2-M2*M2))
280    Y=INT(20*R(I)+0.5)+10
290    IF Y<0 THEN Y=0
300    IF Y>20 THEN Y=20
310    D(Y)=D(Y)+1
320    LINE (200,377)-(620,377),PSET,12
330    LINE (200+20*Y,381-H-H*D(Y))-(219+20*Y,380-H*D(Y)),PSET,3,BF,3
340  NEXT I
350  WHILE L<0.025*K
360    FOR I=1 TO K
370      IF R>=R(I) THEN GOTO 390
380      R=R(I) : J=I
390    NEXT I
400    R0=R
410    R(J)=0 : R=0 : J=0
420    L=L+1
430  WEND
440  LINE (219+20*(20*R0+10),385)-(219+20*(20*R0+10),376),PSET,5
450  PRINT " Sample value of r0(5%)=" ;R0
```

このプログラムを実行して，Trials?に実験回数Kを，Samples?に標本数Nを入力する。続いて，Expectations:e1,e2?に２つの変量の母平均E1とE2を入力し，Standard deviations:s1,s2?にそれらの標準偏差S1とS2を入力する。その結果，期待値がE1，標準偏差がS1である正規母集団とそれに独立な期待値がE2，標準偏差がS2である正規母集団からの大きさNの標本の組みが抽出されて，標本相関係数が計算される。画面上にはこれら全部でK個の標本相関係数の度数分布図がダイナミックに表示され，最後に有意水準5%の両側検定の棄却域の右側の限界値が表示される。この値は標本相関係数に対してシミュレーションで得られた分布から求められたものである。母集団分布として共に一様分布を考える場合にはプログラムの210,220行を215,225行と置き換えれば良い。

このようにしてシミュレーションを活用することによって，理論的には未だ明らかにされていない問題にも実際的で初等的な対処法を見出す可能性のあることが分かる。

4．まとめ

確率の各基本事項を中心に，それらの実際的な理解のためにシミュレーションを活用する方法について述べた。さらに，統計的推測において理論的に導かれた手法に対して，それらを実際に適用する場合の妥当性とともに問題点の理解を，母平均の区間推定法を例としてシミュレーションによって検討した。また，数理的には未だ確率分布が明らかでない統計量であって

も，シミュレーションを用いてその分布を近似的に知ることが出来るので，仮説検定の棄却域を定めるのに利用できることを標本相関係数を例にして示した。

ここで述べた方法は数理に替ってプログラムを書き，問題を直接的にシミュレートすることによって考察するので，実際的で興味深く，予備的な数学の知識も最小限で済まされている。したがって，これから統計的方法を学ぼうとする一般の学生だけでなく，学校で学んだ数学を忘れてしまっている一般社会人にとっても取り組みやすい方法であろうと考えられる。

参考書

[1] 藤曲哲郎・森 隆一・山本 浩 『シミュレーションによる統計学』 1996, 日本評論社